**Stage 1**

# Introduction

In the stage 1 of this project, we used a dataset from the UCI Machine Learning Repository that contained the risk factors of cervical cancer that can be used in two directions, binary classification and multiple classification. The dataset contains 36 attributes and 858 examples. At the beginning of this project, we chose three different classification models: logistic regression, support vector machines (SVM), and neural networks to determine which model has better performance on the validation set.

The original features are as follows:

| Features | | Targets |
| --- | --- | --- |
| **int** | **bool** | **bool** |
| Age | Smokes | Hinselmann |
| Number of sexual partners | Hormonal Contraceptives | Schiller |
| First sexual intercourse(age) | IUD | Cytology |
| Number of pregnancies | STDs | Biopsy |
| Smoke (years) | STDs:condylomatosis |  |
| Smoke (packs/years) | STDs:cervical condylomatosis |  |
| Hormonal Contraceptives (years) | STDs:vaginal condylomatosis |  |
| IUD (years) | STDs:vulvo-perineal condylomatosis |  |
| STDs (number) | STDs:syphilis |  |
| STDs: Number of diagnosis | STDs:pelvic inflammatory disease |  |
| STDs: Time since first diagnosis | STDs:genital herpes |  |
| STDs: Time since last diagnosis | STDs:molluscum contagiosum |  |
|  | STDs:AIDS |  |
|  | STDs:HIV |  |
|  | STDs:Hepatitis B |  |
|  | STDs:HPV |  |
|  | Dx:Cancer |  |
|  | Dx:CIN |  |
|  | Dx:HPV |  |
|  | Dx |  |

# Directions

1. Binary classification

We decided to treat it as a binary classification problem, with the goal of predicting whether a patient will have cervical cancer based on one of the following target variables: Hinselmann, Schiller, Cytology, and Biopsy.

1. Multiple classification

We also decided to consider it as a multiple classification problem, with the goal of predicting the risk degree of a patient will have cervical cancer based on 4 target variables: Hinselmann, Schiller, Cytology, and Biopsy since we found the binary classification cannot improve the accuracy of validation set.

# Data Preprocessing

1. After reviewing all dataset, we found there are too many irrelevant features, incomplete features, and missing values so that we modify them as follows:

* We replaced all ‘?’ with NaN.
* We deleted two irrelevant features which had too many 0s, “Time since first diagnosis” and “Time since second diagnosis”.
* We tried to normalize the integer type features but it didn’t work as we wished so we standardized the dataset since it is more useful in classification than regression.

### Binary classification

Combined four target variables as one named result to represent the true value of whether a patient has cervical cancer or not. Then dropped four target variables: Hinselmann, Schiller, Cytology, and Biopsy.

### Multiple classification

Added four target variables as result, which is df['Biopsy']+df['Citology']+df['Schiller']+df['Hinselmann']

2. We splitted the original data into 80% training and 20% validation.

# **Models –** Binary classification

### Logistic Regression

| Logistic Regression | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | No Regularization | | | Regularization | | | |
| L1 Norm | | L2 Norm | |
| Polynomial Transformation Degree | | | PTD | | PTD | |
| 1 | 2 | | 1 | 2 | 1 | 2 |
| Training | 0.890671 | 0.930029 | | 0.879009 | 0.880466 | 0.883382 | 0.912536 |
| Validation | 0.889535 | 0.854651 | | 0.889535 | 0.889535 | 0.883721 | 0.877907 |

Comparing the performance between polynomial transformation degree equals 2 and no transformation, we know that the performance is not good as we wish so that we tried to add regularization to see whether it will increase the performance of the validation set. However, they are worse than before. We thought about the output for a while. Then we decided not to add regularization and we figured out that we need to add a penalty when the training performance is too high. Obviously, 0.89 is not too high so it is unnecessary to add regularization. For this reason, we didn’t consider adding regularization on the following two models.

### Support Vector Machine

| Support Vector Machine | | |
| --- | --- | --- |
| Dataset | No Regularization | |
| Linear Kernel | Gaussian RBF |
| Training | 0.897959 | 0.889213 |
| Validation | 0.901163 | 0.883721 |

| Polynomial Kernel | | |
| --- | --- | --- |
| Degree | Training | Validation |
| 1 | 0.886297 | 0.895349 |
| 2 | 0.912536 | 0.877907 |
| 3 | 0.921283 | 0.872093 |
| 4 | 0.922741 | 0.860465 |
| 5 | 0.924198 | 0.860465 |
| 6 | 0.925656 | 0.860465 |
| 7 | 0.928571 | 0.860465 |
| 8 | 0.928571 | 0.860465 |

We utilized the linear kernel, polynomial kernel, and gaussian RBF to obtain the performance of training and validation performance. Intuitively, they are similar to the result of using the logistic regression model. Comparing those three techniques, the linear kernel without regularization has the best performance on the validation set.

## Neural Network

| Neural Network | | | |
| --- | --- | --- | --- |
|  | 4 layer Neural Network (30, 30, 30, 1) | 5 layer Neural Network  (30, 30, 30, 30, 1) | 6 layer Neural Network  (30, 30, 30, 30, 30, 1) |
| Epochs | 8 | 8 | 8 |
| Loss | 0.3017 | 0.3109 | 0.3120 |
| Training | 0.895044 | 0.900875 | 0.889213 |
| Validation | 0.901162 | 0.877907 | 0.883721 |

In the neural network model, we tried 3 cases including 4, 5 and 6 layers without regularization. The reason why we chose 8 epochs is because the more epochs, the more overfitting. Then the performance on validation will be worse. In conclusion, the 4 layer Neural network has the best performance on the validation set.

# **Models –** Multiple classification

### Logistic Regression

| Logistic Regression | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | No Regularization | | | Regularization | | | |
| L1 Norm | | L2 Norm | |
| Polynomial Transformation Degree | | | PTD | | PTD | |
| 1 | 2 | | 1 | 2 | 1 | 2 |
| Training | 0.884840 | 0.930029 | | 0.879009 | 0.880466 | 0.883382 | 0.912536 |
| Validation | 0.877907 | 0.854651 | | 0.889535 | 0.889535 | 0.883721 | 0.877907 |

We still compared the performance of the validation set between original and polynomial transformation(d = 2) with regularization or without regularization. The result shows a similar pattern to the binary classification, in which the performance of the validation set is added to the L1 norm without transformation.

### Support Vector Machine

| Support Vector Machine | | |
| --- | --- | --- |
| Dataset | No Regularization | |
| Linear Kernel | Gaussian RBF |
| Training | 0.897959 | 0.880466 |
| Validation | 0.901163 | 0.883721 |

| Polynomial Kernel | | |
| --- | --- | --- |
| Degree | Training | Validation |
| 1 | 0.884840 | 0.883721 |
| 2 | 0.900875 | 0.872093 |
| 3 | 0.905248 | 0.877907 |
| 4 | 0.905248 | 0.877907 |
| 5 | 0.905248 | 0.877907 |
| 6 | 0.905248 | 0.872093 |
| 7 | 0.905248 | 0.872093 |
| 8 | 0.906706 | 0.872093 |
| 9 | 0.906706 | 0.872093 |

We still utilized the linear kernel, polynomial kernel, and gaussian RBF to obtain the performance of training and validation performance. Intuitively, they are similar to the result of using the logistic regression model. Comparing those three techniques, the linear kernel without regularization has the best performance on the validation set.

### Neural Network

| Neural Network | | | |
| --- | --- | --- | --- |
|  | 4 layer Neural Network (30, 30, 30, 1) | 5 layer Neural Network  (30, 30, 30, 30, 1) | 6 layer Neural Network  (30, 30, 30, 30, 30, 1) |
| Epoches | 8 | 8 | 8 |
| Loss | 0.4055 | 0.4569 | 0.4757 |
| Training | 0.810496 | 0.848396 | 0.879009 |
| Validation | 0.813953 | 0.819767 | 0.889535 |

In the neural network model, we tried 3 cases including 4, 5 and 6 layers without regularization. The reason why we chose 8 epochs is because the more epochs, the more overfitting. Then the performance on validation will be worse. In sum, the 6 layer Neural network has the best performance on the validation set.

# Note：

After we finished the stage1, we found that no matter how we changed the model's parameters, the performance of the validation set still kept the same, so we intend to make some changes like paying more attention to the feature itself. According to the article, “the HPV virus plays a crucial impact in cervical cancer”, Cervical cancer occurs in a woman's cervix (the entrance to the uterus from the vagina). Most cases of cervical cancer (99%) are associated with high-risk human papillomavirus (HPV) infection, a widespread virus transmitted through sexual contact. While most HPV infections resolve spontaneously and do not cause symptoms, persistent infection can lead to cervical cancer. Cervical cancer is the fourth most common cancer in women(WHO,1).

**Stage 2**

# Introduction

In the stage 2 of this project, we properly modified our dataset to be easier to handle according to studying the cause of cervical cancer in depth in our subsequent study to help the model calculate better performance. For example, we hope patients know their situation without doing the redundant tests.

# Directions

1. Binary classification

We tried to predict patients' Biopsy results according to a modified dataset.

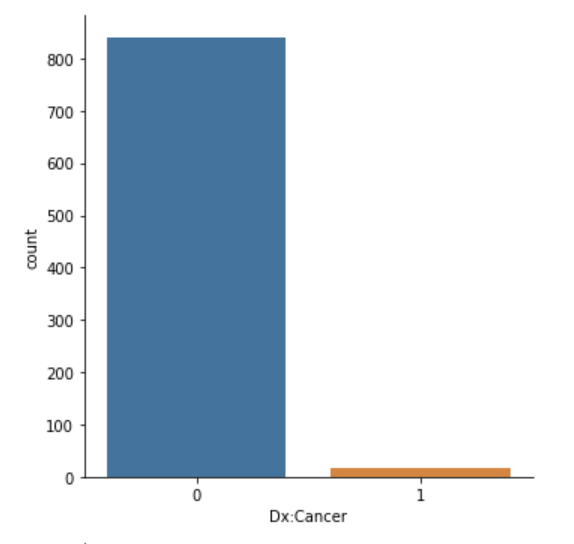
1. Multiple classification

We also decided to consider it as a multiple classification problem, with the goal of predicting the risk degree of a patient will have cervical cancer based on 4 target variables: Hinselmann, Schiller, Cytology, and Biopsy.

# Data Preprocessing

1. After dealing with all features, we remodified them as follows:

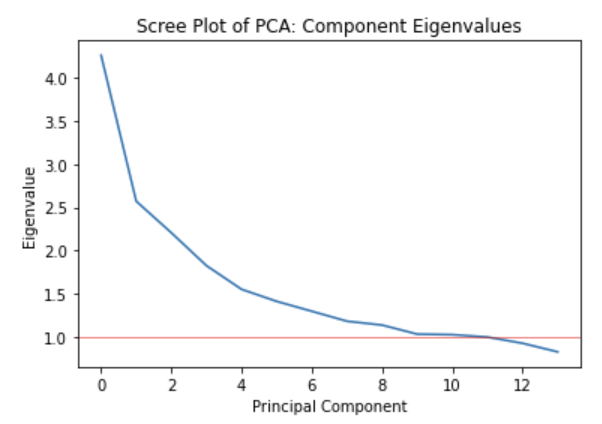
* Replaced all ‘?’ values with NaN.
* Find NaN values’ numbers in each feature.
* Find imbalance data.



1. For those bool type NaN data, we modified it as 1 according to our research in medical psychology.

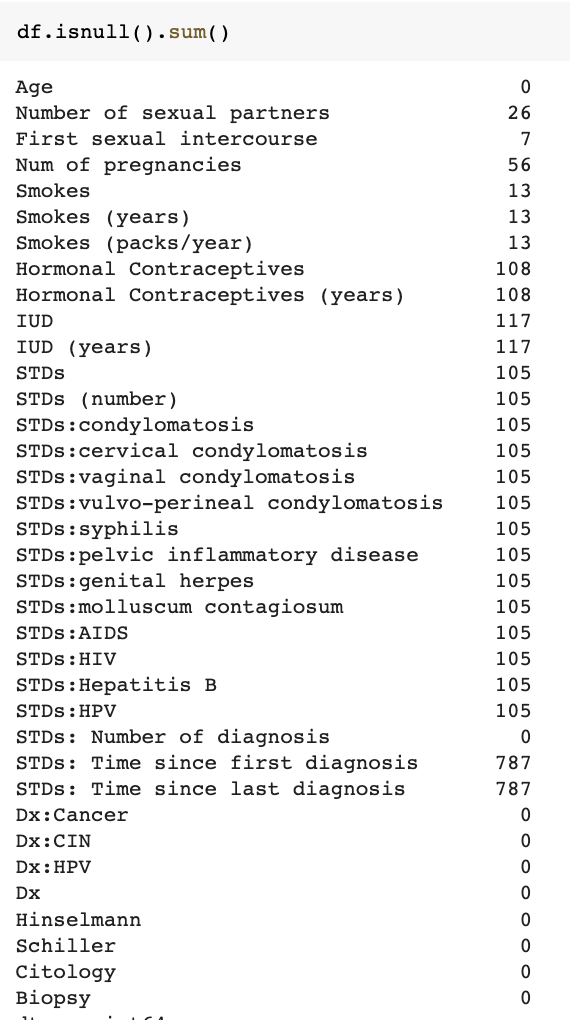
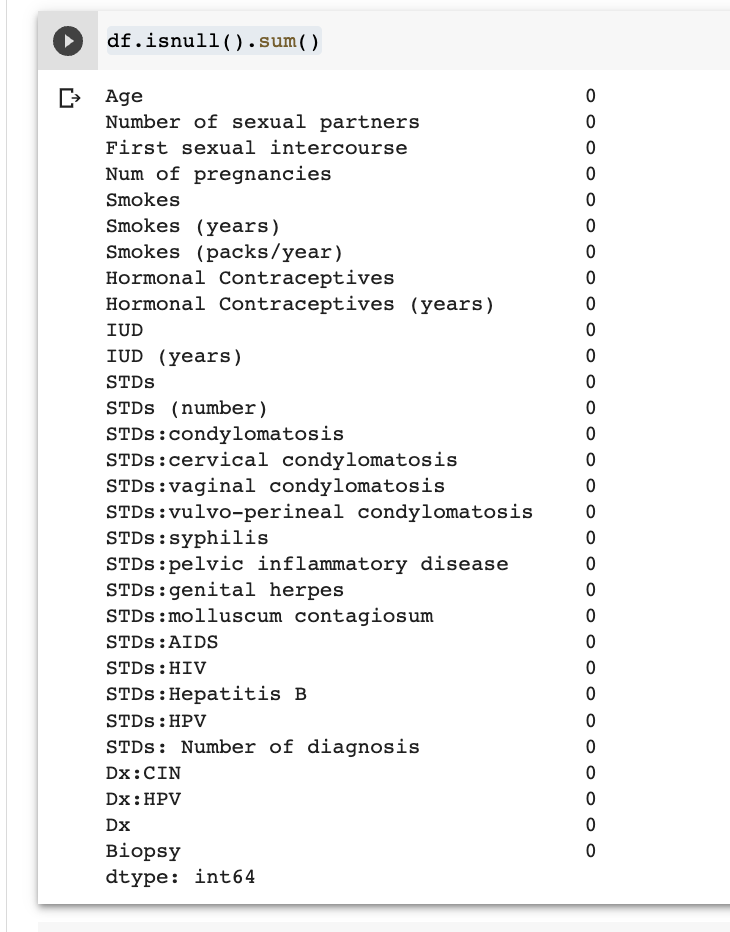
* For some binary examples of NAN data, we added 1 to the data of these NANs. Because, according to our speculation, for these negative things, most patients do not want others to know their privacy, such as whether they have HIV (HIV is discriminated against and prejudiced in society), whether they smoke, and whether they have a contraceptive ring installed. Therefore, we treated unanswered questions as 1.

1. We will fill in the median to allow the data to be processed by the model for the rest of the empty data.
2. We used the PCA algorithm to reduce the number of features due to excessive features after a specific binary or multiple classification process.



1. We splitted the original data into 80% training and 20% validation.

1. Finally, we check the complicity of the dataset.

### Binary classification

### Removed other target variables ('Hinselmann,' 'Schiller,' 'Cytology'). We should process the data better and reasonably rather than just pursuing the correct rate and obscuring what needs to be studied. After investigation, we found that Biopsy had the highest accuracy in determining cervical cancer, so we clarified our target by choosing Biopsy as the only target variable and removing other target variables ('Hinselmann,' 'Schiller,' 'Cytology'), and we hoped that by Asking patients about their habits, medical history, and personal information, we hope to assess patients' risk of disease quickly.

1. Multiple classification：

Added four target variables as result, which is df['Biopsy']+df['Citology']+df['Schiller']+df['Hinselmann']

# Models

| Dataset | Binary Classification | | Multiple Classification | |
| --- | --- | --- | --- | --- |
| Training | Validation | Training | Validation |
| Logistic Regression | 0.941667 | 0.922481 | 0.880466 | 0.872093 |
| SVM | 0.941667 | 0.926357 | 0.886297 | 0.889535 |
| Neural Network | 0.943333 | 0.926357 | 0.921282 | 0.872093 |

### In sum, we found the best performance of predicting the result of biopsy is the SVM and Neural Network model, which is 92.6%. The reason why we didn’t add the regularization part is because our output is not overfitting.

# Conclusion

We found that the accuracy was greatly improved with the same model (logistic regression, SVM, neural network) after a more refined processing of the data. This also shows that for these unfamiliar data, we should first understand the meaning behind these data, understand cervical cancer from the medical perspective, understand the characteristics and relationship of each feature, and then perform targeted processing on the features to really improve the prediction accuracy.

Our model can be applied to the early prediction of cervical cancer, which allows patients to get the biopsy prediction of cervical cancer by filling out a questionnaire without intervening any clinical examination, with an accuracy of 92-93%, which allows patients to predict their condition earlier and go to the hospital for further examination in time

Through this project, we found a rule that it is easier to get better performance with adequate and comprehensive data than with model selection.

# Works Cited

Machine Learning. Imbalanced Data. 2022.

<https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data>

Accessed 7 Dec. 2022.

Zakaria Jaadi. A Step-by-Step Explanation of Principal Component Analysis (PCA). 2022.

<https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

Accessed 7 Dec. 2022.

WHO's view, the HPV virus plays a crucial impact in cervical cancer. WHO. cervical-cancer, 2022. <https://www.who.int/health-topics/cervical-cancer>

Accessed 7 Dec. 2022.